

# ST502: Final Project

Lamia Benyamine

2024-12-09

## Introduction to McNemar's test

McNemar's test is a statistical test for dependent (paired) categorical data. McNemar's is best used for data with matched pairs, for example, comparing pre-treatment with post-treatment.

## Analyzing the Dataset

To test a restricted multinomial vs a free multinomial, we have:

- $H_0$ : No relationship between drug and relief, or,  $\pi_{1\bullet} = \pi_{\bullet 1}$  and  $\pi_{2\bullet} = \pi_{\bullet 2}$  equivalent to  $\pi_{12} = \pi_{21}$   
 $H_A$ : cell probabilities are 'free' (other than the sum to 1 constraint)

$$\sum_{i=1}^I \sum_{j=1}^J \pi_{ij} = 1$$

- Our test statistic is given as Pearson's Chi-Square test statistic

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

and a reference distribution of  $\chi^2$ .

- For the Rejection region and p-value, we use R and assume  $\alpha = 0.05$ .

```
cuttoff <- qchisq(0.95, df = 1)
LRT <- 2*sum(15*log(15/40))
```

- So the RR = { xobs > 3.84145882069412} and p-value = P(  $\chi^2$  RV  $\geq$  -29.4248775903518)

Create the data matrix and run McNemar's test to determine if we have evidence that the reflux drugs have different probabilities of relief.

```
reflux_data <- matrix(c(85,15, 40,110), nrow = 2, ncol = 2, byrow = TRUE,
                      dimnames = list(c("A-Success", "A-Failure"),
                                       c("B-Success", "B-Failure")))
reflux_data
```

```
##           B-Success B-Failure
## A-Success      85       15
## A-Failure      40      110
```

```
#run the test for the given data
robs <-mcnemar.test(reflux_data, correct = FALSE)
robs
```

```
##
## McNemar's Chi-squared test
##
## data:  reflux_data
## McNemar's chi-squared = 11.364, df = 1, p-value = 0.000749
```

Based on the McNemar's output, the p-value is low and  $11.3636363636364 > 3.84145882069412$ , so we reject the null in favor of the alternative. We conclude that there is sufficient evidence of an association between drug type and reflux relief.

## Derive Parts of the Test

1. Show this is true  $\pi_{1\bullet} = \pi_{\bullet 1}$  and  $\pi_{2\bullet} = \pi_{\bullet 2}$  equivalent to  $\pi_{12} = \pi_{21}$

$\pi_{1\bullet}$  is the sum of all the probabilities in row 1. So  $\pi_{1\bullet} = \pi_{11} + \pi_{12}$  and similarly  $\pi_{2\bullet} = \pi_{21} + \pi_{22}$

$\pi_{\bullet 1}$  is the sum of all the probabilities in column 1. So  $\pi_{\bullet 1} = \pi_{11} + \pi_{21}$  and similarly

$\pi_{\bullet 2} = \pi_{12} + \pi_{22}$

Given,  $\pi_{1\bullet} = \pi_{\bullet 1}$  and  $\pi_{2\bullet} = \pi_{\bullet 2}$ , we substitute and get:

$$\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21} \quad (1)$$

$$\pi_{21} + \pi_{22} = \pi_{12} + \pi_{22} \quad (2)$$

To simplify we subtract  $\pi_{11}$  from both sides in (1), and subtract  $\pi_{22}$  from both sides in (2).

$$\pi_{12} = \pi_{21} \quad (1)$$

$$\pi_{21} = \pi_{12} \quad (2)$$

Therefore the statement in the null hypothesis is proven true.

2. Second, under this null restriction on our multinomial, derive the maximum's for  $\pi_{11}, \pi_{12}, \pi_{21}$ , and  $\pi_{22}$ . This can be done using Lagrange multipliers or by substituting in carefully to include the restriction. Starting with our general likelihood

$$L(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) \propto c \prod_{i=1}^I \prod_{j=1}^J \pi_{ij}^{n_{ij}}$$

The log-likelihood is then

$$l(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = c + n_{11} \ln(\pi_{11}) + n_{12} \ln(\pi_{12}) + n_{21} \ln(\pi_{21}) + n_{22} \ln(\pi_{22})$$

Using Lagrange Multipliers, derive the maximums subject to the sum to 1 constraints. Since the constraints can be rewritten as  $\sum_{i=1}^2 \pi_{i\bullet} - 1 = 0$  and  $\sum_{j=1}^2 \pi_{\bullet j} - 1 = 0$ , we add these constraints into our log-likelihood with Lagrange multipliers.

$$l(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = c + n_{11} \ln(\pi_{11}) + n_{12} \ln(\pi_{12}) + n_{21} \ln(\pi_{21}) + n_{22} \ln(\pi_{22}) + \lambda \left( \sum_{i=1}^2 \sum_{j=1}^2 \pi_{ij} - 1 \right)$$

We get the partial derivatives for all variables, set to 0, and solve.

$$\frac{\partial l}{\partial \pi_{11}} = \frac{n_{11}}{\pi_{11}} + \lambda \equiv 0 \Leftrightarrow \pi_{11} = -\frac{n_{11}}{\lambda} \quad (1)$$

$$\frac{\partial l}{\partial \pi_{12}} = \frac{n_{12}}{\pi_{12}} + \lambda \equiv 0 \Leftrightarrow \pi_{12} = -\frac{n_{12}}{\lambda} \quad (2)$$

$$\frac{\partial l}{\partial \pi_{21}} = \frac{n_{21}}{\pi_{21}} + \lambda \equiv 0 \Leftrightarrow \pi_{21} = -\frac{n_{21}}{\lambda} \quad (3)$$

$$\frac{\partial l}{\partial \pi_{22}} = \frac{n_{22}}{\pi_{22}} + \lambda \equiv 0 \Leftrightarrow \pi_{22} = -\frac{n_{22}}{\lambda} \quad (4)$$

Use the sum to 1 constraint to solve for  $\lambda$ .

$$\begin{aligned} \pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1 &\Leftrightarrow -\frac{n_{11}}{\lambda} - \frac{n_{12}}{\lambda} - \frac{n_{21}}{\lambda} - \frac{n_{22}}{\lambda} = 1 \\ &\Rightarrow \lambda = -(n_{11} + n_{12} + n_{21} + n_{22}) = -n \end{aligned}$$

Substituting  $\lambda$  in equations 1-4 above, we get the below critical values.

$$\pi_{11} = \frac{n_{11}}{n} \quad (1)$$

$$\pi_{12} = \frac{n_{12}}{n} \quad (2)$$

$$\pi_{21} = \frac{n_{21}}{n} \quad (3)$$

$$\pi_{22} = \frac{n_{22}}{n} \quad (4)$$

3. Derive the form of the LRT for a restricted vs free multinomial for this specific problem. Show that

$$-2\ln \left( \frac{L(\pi_{11}^{\sim}, \pi_{12}^{\sim}, \pi_{21}^{\sim}, \pi_{22}^{\sim})}{L(\pi_{11}^{\wedge}, \pi_{12}^{\wedge}, \pi_{21}^{\wedge}, \pi_{22}^{\wedge})} \right) = 2 \sum_{i=1}^2 \sum_{j=1}^2 Obs_{ij} \ln \left( \frac{Obs_{ij}}{Exp_{ij}} \right)$$

and then argue that the appropriate reference distribution has 1 degree of freedom.

We know our likelihood is a multinomial

Max over  $\omega_0$  : we know  $\pi_{ij}^{\wedge} = \frac{n_{ij}}{n}$  from above.

Max over  $\Omega$  :  $(\pi_{ij}^{\wedge})_{\sim} = \frac{n_{i\bullet} * n_{\bullet j}}{n}$

Observed LRT Statistic:

$$\begin{aligned} \Lambda &= \left( \frac{L(\pi_{11}^{\sim}, \pi_{12}^{\sim}, \pi_{21}^{\sim}, \pi_{22}^{\sim})}{L(\pi_{11}^{\wedge}, \pi_{12}^{\wedge}, \pi_{21}^{\wedge}, \pi_{22}^{\wedge})} \right) = \frac{\prod_{i=1}^I \prod_{j=1}^J (\pi_{i\bullet}^{\sim} \pi_{\bullet j}^{\sim})^{n_{ij}}}{\prod_{i=1}^I \prod_{j=1}^J (\pi_{ij}^{\wedge})^{n_{ij}}} \\ &\Rightarrow \prod_{i=1}^I \prod_{j=1}^J \left( \frac{(\pi_{i\bullet}^{\sim} \pi_{\bullet j}^{\sim})}{\pi_{ij}^{\wedge}} \right)^{n_{ij}} \end{aligned}$$

By our large-sample theory

$$-2\ln(\Lambda) \sim^{H_0} \chi^2_{m-k-1}$$

Rewriting our test statistic and substituting in known values, we get

$$\begin{aligned} -2\ln(\Lambda) &= -2 \sum_{i=1}^I \sum_{j=1}^J n_{i,j} \ln \left( \frac{(\pi_{i\bullet}^{\sim} \pi_{\bullet j}^{\sim})}{\pi_{ij}^{\wedge}} \right) \Rightarrow 2 \sum_{i=1}^I \sum_{j=1}^J n_{i,j} \ln \left( \frac{(\pi_{i\bullet}^{\sim} \pi_{\bullet j}^{\sim})}{\pi_{ij}^{\wedge}} \right) \\ &\Rightarrow 2 \sum_{i=1}^I \sum_{j=1}^J n_{i,j} \ln \left( \frac{\frac{n_{ij}}{n_{i\bullet} * n_{\bullet j}}}{\frac{n_{ij}}{n}} \right) \Rightarrow 2 \sum_{i=1}^I \sum_{j=1}^J n_{i,j} \ln \left( \frac{n_{ij}}{n} \right) \end{aligned}$$

We know  $n_{i,j}$  is the observed counts and from the derivation above  $\frac{n_{ij}}{n}$  is the expected counts under  $H_0$ . So  $n_{i,j} = Obs_{ij}$  and  $\frac{n_{ij}}{n} = Exp_{ij}$ . Now substituting these into our test statistic we get:

$$-2\ln(\Lambda) = 2 \sum_{i=1}^2 \sum_{j=1}^2 Obs_{ij} \ln \left( \frac{Obs_{ij}}{Exp_{ij}} \right)$$

To determine degrees of freedom  $(m - k - 1)$ , we need to determine the dimensions of  $\omega_0$  and  $\Omega$ .

For  $H_0$  :  $\dim \omega_0 = \{ (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) : \pi_{12} = \pi_{21} \} = 2 = k$

For  $H_A$  :  $\dim \Omega = \{ (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) : \text{free other than the sum to 1 constraint} \} = 4 = m$

$df = 4 - 2 - 1 = 1$

So the reference distribution has 1 degree of freedom.

4. Lastly, we know we can use Pearson's chi-square test statistic instead of this LRT and they are asymptotically equivalent. In our above data example, we used

$$\chi^2 = 2 \sum_{i=1}^2 \sum_{j=1}^2 \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Show that Pearson's chi-square test statistic can be simplified into the form on the right.

From above we know,  $n_{i,j} = Obs_{ij}$  and  $\frac{n_{ij}}{n} = Exp_{ij}$ . Under  $H_0$ , with the assumption that  $\pi_{12} = \pi_{21}$ , we can combine the terms and have  $\pi_{12} = \pi_{21} = \frac{n_{12} + n_{21}}{2n}$ . Since we are dealing with discordant pairs, we know  $n_{11} = n_{22} = 0$ .

So substituting these values into Pearson's chi-square test statistic we get:

$$\begin{aligned} 2 \sum_{i=1}^2 \sum_{j=1}^2 \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} &= 2 \left[ \frac{(n_{12} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/4} + \frac{(n_{21} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/4} \right] \\ &= \left[ \frac{((n_{12} - (n_{12} + n_{21})/2))^2}{(n_{12} + n_{21})/2} + \frac{((n_{21} - (n_{12} + n_{21})/2))^2}{(n_{12} + n_{21})/2} \right] = \frac{(\frac{n_{12} - n_{21}}{2})^2 + (\frac{n_{21} - n_{12}}{2})^2}{(n_{12} + n_{21})/2} \\ &= \frac{(\frac{n_{12}^2}{4} - \frac{n_{12}n_{21}}{2} + \frac{n_{21}^2}{4}) + (\frac{n_{12}^2}{4} - \frac{n_{12}n_{21}}{2} + \frac{n_{21}^2}{4})}{(n_{12} + n_{21})/2} = \frac{(\frac{n_{12}^2}{2} - n_{12}n_{21} + \frac{n_{21}^2}{2})}{(n_{12} + n_{21})/2} \\ &= \frac{(n_{12}^2 - 2n_{12}n_{21} + n_{21}^2)}{n_{12} + n_{21}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \end{aligned}$$

Thus we have shown we can simplify the test statistic to the requested form.

## Simulation Study

Using simulation based methods is an easier way to find approximate results for the power of a test. We'll investigate the  $\alpha$  of the Pearson chi-square test and its power.

## McNemar's Function

Create a function to conduct McNemar's test and return True if there is enough evidence to reject  $H_0$  and False otherwise.

Kept getting this error '*Correlations are beyond their upper limits imposed by expectations*', so I added a condition to check the correlation and adjust it if it is higher than the maximum allowed correlation.

```
library(MultiRNG)
mcnemar_func <- function(n, p, pi1, pi2){

  #check if correlation is within range, if not, sub in max correlation value
  max_p <- sqrt(pi1 * (1 - pi1) * pi2 * (1 - pi2))
  if (abs(p) > max_p) {
    p <- max_p}

  #store parameter values in a vector/matrix
  propvec <- c(pi1, pi2)
  cmat <- matrix(c(1, p, p, 1), nrow=2, ncol=2)

  #generate correlated binary data
  bin_data <- draw.correlated.binary(no.row = n, d = 2, prop.vec = propvec, corr.mat = cmat)

  #use a contingency table to find observed counts
  obs_count <- table(bin_data[,1], bin_data[,2])

  #check if there is a row for each scenario, if not create one with value 0
  row <-rownames(obs_count)
  col <- colnames(obs_count)
  n_12 <- ifelse("0" %in% row && "1" %in% col, obs_count["0", "1"], 0)
  n_21 <- ifelse("1" %in% row && "0" %in% col, obs_count["1", "0"], 0)

  #return FALSE if there are no discordant pairs (to avoid NAs)
  if (n_12 == 0 && n_21 == 0) {
    return(FALSE)}

  #use Pearson's Chi-square test stat
  xobs <- ((n_12 - n_21)^2)/(n_12 + n_21)

  #determine the cut off value of failing to reject null hypothesis
  cutoff <- qchisq(0.95, df = 1)

  #return True for Reject H0 and False for Failing to Reject H0
  return(xobs > cutoff)
}
```

## Generate combinations of data

- We'll generate data under all combinations of the following:

```

#sample size
n_val <- c(25, 40, 80, 200)
#‘drug A’ variable’s success probability
pi1_val <- c(0.1, 0.4, 0.8)
#‘drug B’ variable’s success probability
pi2_val <- c(pi1_val, pi1_val + 0.02, pi1_val + 0.05, pi1_val + 0.1)
#this is the correlation with which we generate our data
p_val <- c(0, 0.2, 0.5)

```

- Create a function to generate the data for all combinations of the parameters

```

get_data <- function(n_val, p_val, pi1_val, pi2_val, N = 100) {
  #set the seed for the random num generation with replicate
  set.seed(25)
  #initialize a list to store data
  gen_data <- list()

  #generate data by looping through all combinations
  for (n in n_val) {
    for (p in p_val) {
      for (pi1 in pi1_val) {
        for (pi2 in pi2_val) {
          #create column names for each combination
          col_name <- paste0("n", n, "_corr", p, "_pi1", pi1, "_pi2", pi2)
          #replicate the test and store data in the list
          gen_data[[col_name]] <- replicate(N,
                                            {mcnemar_func(n = n, p = p, pi1 = pi1, pi2 = pi2)})
        }
      }
    }
  }
  return(gen_data)
}

```

### Replicate the data

We'll generate  $N = 1000$  datasets under each of these settings. We'll be able to determine  $\alpha$  control by looking at the case when  $\pi_1 = \pi_2$ . All of the other cases will allow us to investigate power under the alternative created by the difference in  $\pi$ 's and correlation.

```

samples <- get_data(n_val = n_val, p_val = p_val,
                    pi1_val = pi1_val, pi2_val = pi2_val, N = 1000)

```

### Find Power

Find the probability of rejecting  $H_0$  for each simulated combination.

```

#apply the mean to each list to get the power
power <- lapply(X = samples, FUN = mean)

#create a data frame to better visualize the data

```

```

power_df <- data.frame(
  Combination = names(samples),
  Power = unlist(power) #simplify values to a vector
)

#split the combination column into a vector with 4 values for each parameter
split_combos <- strsplit(x = as.character(power_df$Combination), split = "_")

#extract each parameter value and assign it to a new column
## remove characters from the values
power_df$n <- sapply(split_combos, function(x) x[1])
power_df$n <- sub("n", "", power_df$n)

power_df$Correlation <- sapply(split_combos, function(x) x[2])
power_df$Correlation <- sub("corr", "", power_df$Correlation)

power_df$pi1 <- sapply(split_combos, function(x) x[3])
power_df$pi1 <- sub("pi1", "", power_df$pi1)

power_df$pi2 <- sapply(split_combos, function(x) x[4])
power_df$pi2 <- sub("pi2", "", power_df$pi2)

#remove row names for a cleaner table since values are
#already available in Combination column
row.names(power_df) <- NULL

#set data structure of columns
power_df$pi1 <- as.numeric(power_df$pi1)
power_df$pi2 <- as.numeric(power_df$pi2)
power_df$Correlation <- as.factor(power_df$Correlation)
power_df$n <- as.numeric(power_df$n)

#create a new column pi2-pi1 to use when creating plots
power_df$pi2_pi1 <- (power_df$pi2 - power_df$pi1)

```

## Plots

Read in packages to create better visualizations than with base R.

```

library(ggplot2)
library(tidyr)
library(dplyr)

```

Create Power Plots for each instance of  $\pi_1$  to determine how well the asymptotic rejection region performs at controlling  $\alpha$  and the power of the asymptotic test when comparing certain alternatives.

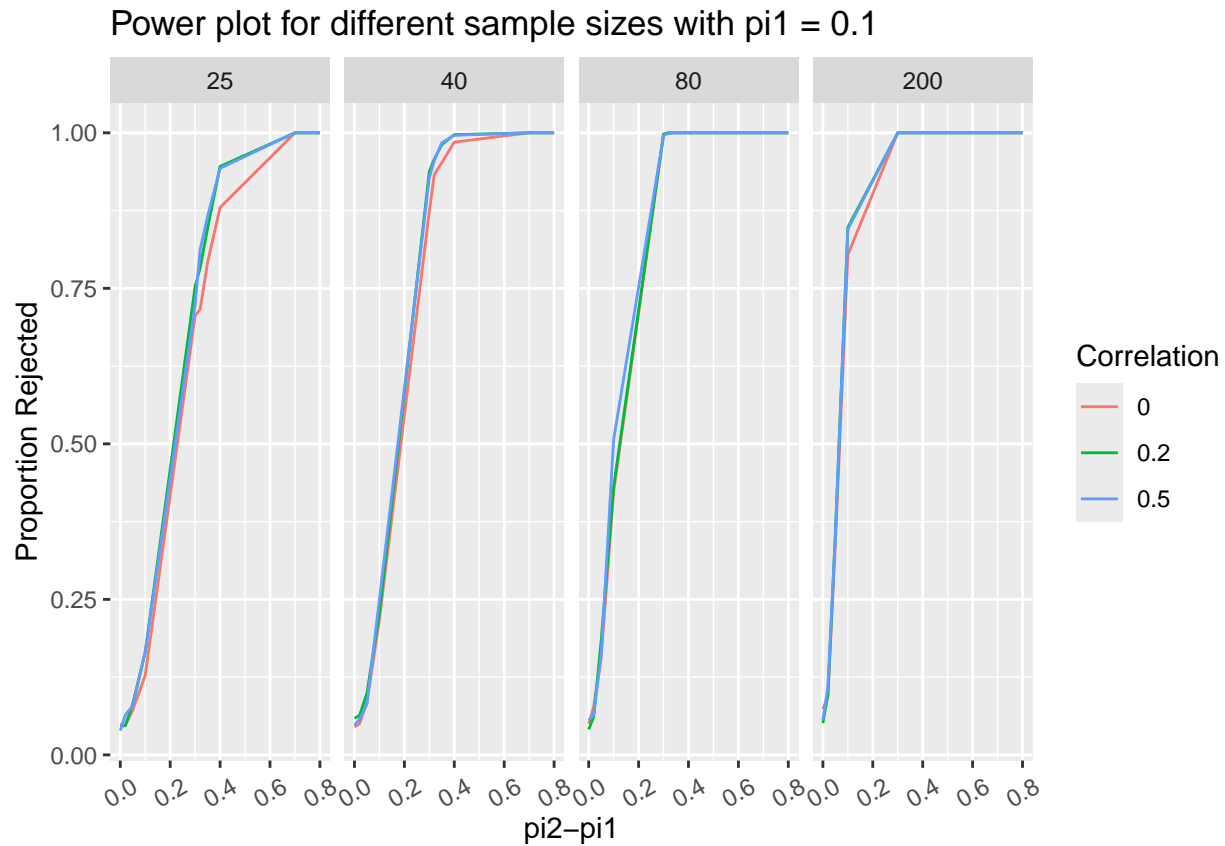
```

power_df <- as_tibble(power_df)

#pi=0.1
power_df |>
  filter(power_df$pi1 == 0.1) |>
  ggplot(aes(x = pi2_pi1, y = Power, group = Correlation, color = Correlation)) +

```

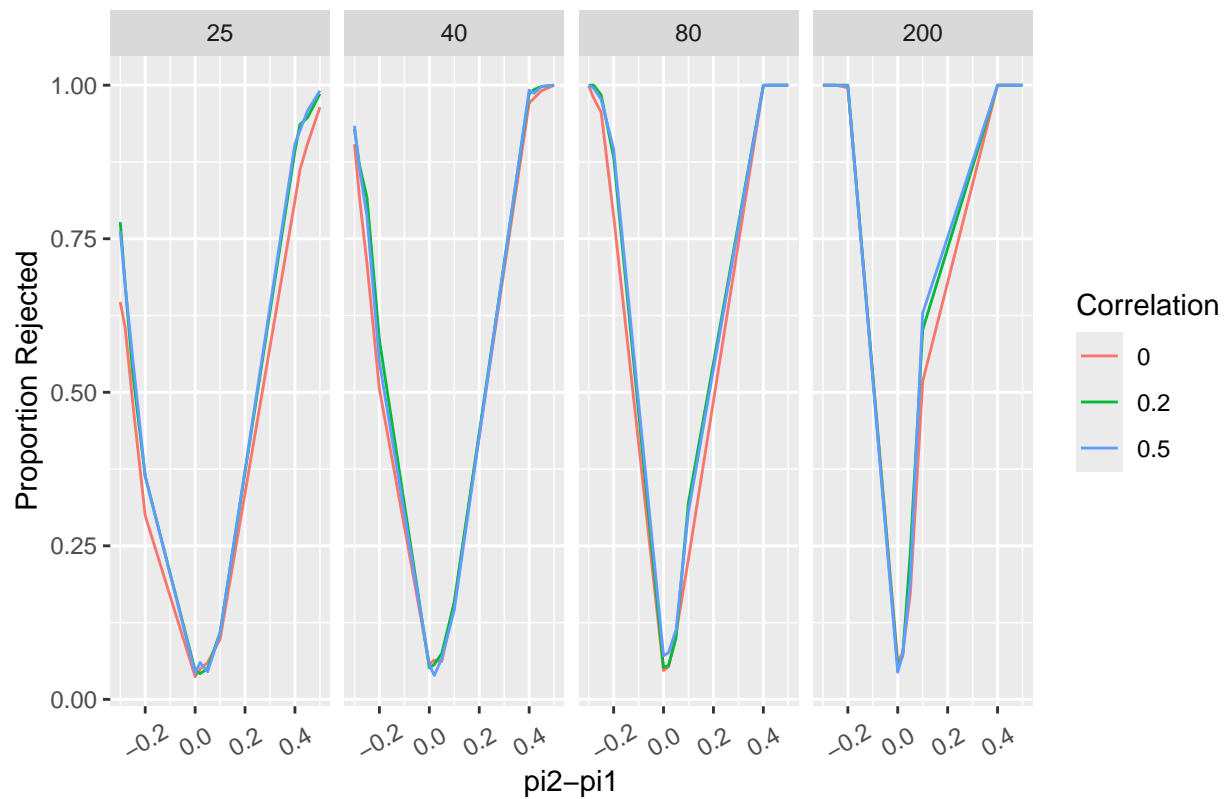
```
geom_line() +
labs(title = "Power plot for different sample sizes with pi1 = 0.1",
      x = "pi2-pi1", y = "Proportion Rejected") +
facet_grid(cols = vars(n)) +
theme(axis.text.x = element_text(angle = 30, vjust = 0.5, hjust=0.5))
```



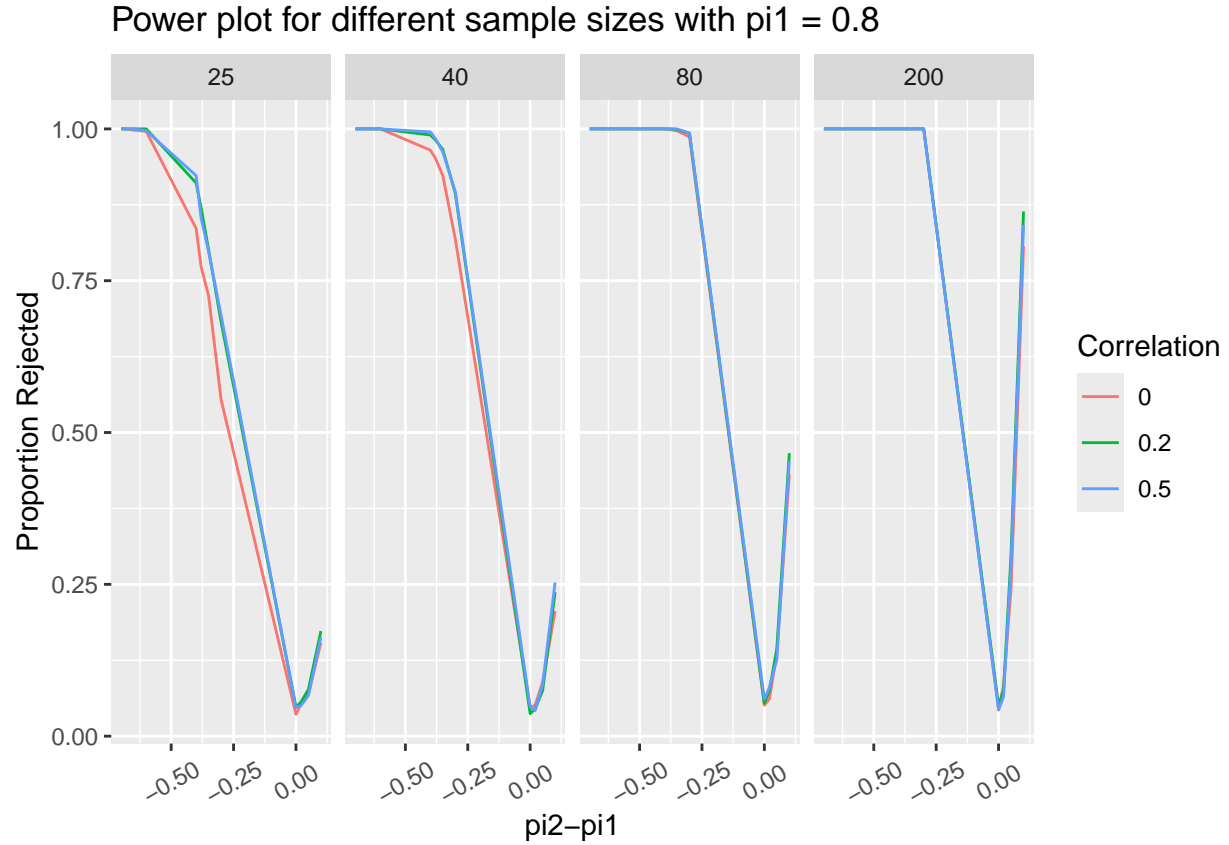
```
#pi=0.4
power_df|>
  filter(power_df$pi1 == 0.4) |>
  ggplot(aes(x = pi2_pi1, y = Power, group = Correlation, color = Correlation)) +
  geom_line() +
  labs(title = "Power plot for different sample sizes with pi1 = 0.4",
        x = "pi2-pi1", y = "Proportion Rejected") +
  facet_grid(cols = vars(n)) +
  theme(axis.text.x = element_text(angle = 30, vjust = 0.5, hjust=0.5))
```



Power plot for different sample sizes with  $\pi_1 = 0.4$



```
#pi=0.8
power_df|>
  filter(power_df$pi1 == 0.8) |>
  ggplot(aes(x = pi2_pi1, y = Power, group = Correlation, color = Correlation)) +
  geom_line() +
  labs(title = "Power plot for different sample sizes with pi1 = 0.8",
        x = "pi2-pi1", y = "Proportion Rejected") +
  facet_grid(cols = vars(n)) +
  theme(axis.text.x = element_text(angle = 30, vjust = 0.5, hjust=0.5))
```



### Results

In all values of  $\pi_1$  where  $\pi_1 = \pi_2$ , the proportion of tests rejecting  $H_0$  was the closest to 0. As the  $|\pi_2 - \pi_1|$  increases, the proportion of tests rejected also increases. The various correlation values acted very similarly, and as the sample sizes increases, the proportion of tests rejected reached high values sooner.